The nucleotide sequence of the nitrogen-regulation gene ntrA of Klebsiella pneumoniae and comparison with conserved features in bacterial RNA polymerase sigma factors

M.J.Merrick and J.R.Gibbins

AFRC Unit of Nitrogen Fixation, University of Sussex, Brighton BN1 9RO, UK

Received 6 September 1985; Accepted 8 October 1985

### ABSTRACT

The nucleotide sequence of the <u>Klebsiella pneumoniae</u> ntragene has been determined. Ntragence a 53,926 Dalton acidic polypeptide; a calculated molecular weight which is significantly lower than that determined by SDS polyacrylamide gel analysis. Ntragence is followed by another open-reading frame (orf) of at least 75 amino acids. In the spacer region between ntragence and orf there are no apparent transcription termination or promoter sequences and therefore orf may be co-transcribed with ntragence.

Previous authors have proposed that NtrA could act as an RNA polymerase sigma factor but the NtrA amino acid sequence does not show a high level of homology to any known sigma factor. However analysis of sequences of five sigma factors from E. coli and B. subtilis has identified two conserved sequences at the C-terminal end of all these polypeptides. These sequences resemble those found in known site-specific DNA-binding domains and may be involved in recognition of conserved -35 and -10 promoter sequences. A similar pair of sequences is present at the C-terminus of NtrA and could play a role in recognition of ntr-activatable promoters.

## INTRODUCTION

In <u>Klebsiella pneumoniae</u> expression of the nitrogen fixation (<u>nif</u>) genes and a number of other genes involved in nitrogen assimilation (e.g. <u>glnA</u>) is regulated by the nitrogen regulation (<u>ntr</u>) system which comprises three genes, <u>ntrA</u>, <u>ntrB</u> and <u>ntrC</u> (1-7). The <u>ntrBC</u> genes are part of a complex operon <u>glnA-ntrBC</u> (8,9) and the <u>ntrA</u> gene is linked to <u>argG</u> (1). A homologous nitrogen control system is present in <u>Salmonella typhimurium</u> (10,11), <u>Escherichia coli</u> (12,13), and <u>Klebsiella aerogenes</u> (14) but in <u>E. coli</u> and <u>K. aerogenes</u> <u>ntrA</u>, <u>B</u> and <u>C</u> are designated glnF, glnL and glnG respectively.

The ntrA product (NtrA) is required together with either

# MATERIALS AND METHODS

# Cloning and DNA sequencing

Restriction enzymes and DNA-modifying enzymes were obtained from commercial sources and used according to the manufacturers' The sequencing strategy was based on a detailed instructions. restriction map of the 1.9 kb ClaI fragment carrying ntrA which had been determined previously (7). Dideoxy sequencing reactions were carried out using defined restriction fragments cloned in M13 mp8, mp9 and mp11 vectors (25) with  $[\alpha^{35}S]$ -dATP as the labelled nucleotide (26). Starting material for construction of M13 clones was derived from plasmids pMM17 and pBS1 (7) and pMM29 (Fig. 1). Plasmid pMM29 carries a 359 bp EcoRI-PvuII fragment from pMM17 cloned between the EcoRI and SmaI sites of the translational fusion vector pMC1403 (27). resultant ntrA-lacZ fusion is in frame as judged by expression of  $\beta$ -galactosidase.

Computer analysis of sequence data employed programs developed by R. Staden (28,29).

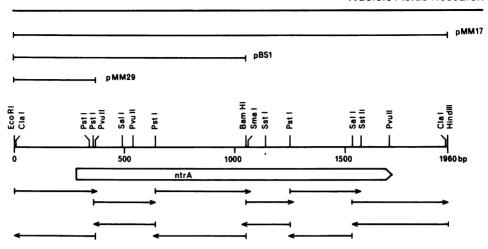


Fig. 1. Restriction map and nucleotide sequencing strategy for the <u>ClaI</u> fragment carrying <u>ntrA</u> from pMM17. The extent of fragments in two derivative <u>ntrA-lacZ</u> translational fusion plasmids pMM29 and pBS1, is also shown. Arrows below the map show the extent of sequence determined from each M13 clone.

## RESULTS

# The nucleotide sequence of ntrA

The sequence of the 1966 nucleotide <a href="EcoRI-HindIII"><u>EcoRI-HindIII</u></a> fragment from pMM17 is shown in Fig. 2. This sequence was determined with contiguous overlapping readings for both strands and any single base position averaged more than 5 readings.

The direction of transcription of <a href="https://ntmailto.com/ntmai

In order to confirm that the gene does terminate 250 bp before the  $\underline{\text{Cla}}\text{I}$  site, plasmid pMM17 was linearised by

3777.	10	31 1 1 G	20	INICA	30	MICA	40	INII	50	ccigc	60	ICAGCG	70		80	AACCGACCA 90	
																	_
TAACG		GAAAC		CGTTT		CGTG		CGTC		GGGCC		SATAGE		TACG		GCAAATCCT	
	100		110		120		130		140		150		160		170	180	
TGAAGA	190	CAGGT		CGTGI	210	reces.		CAGA		ATAGG		AGGTTA	250	1-1-1-1-7	260	GAGATATTO 270	!
	190		200		210		220		230		240		250		260	2/0	
		M K					s Q	_							I R	LLQ	27
cccom																CTACTGCAG	
acce 1	280	IGAAG	290	TIGG	300	100011	310	CAGC	320	ATGAC	330	AACIG	340	GCGA.	350	360	
	200		290		300		310		320		330		340		330	300	
L S	T 1	LE	L O	^ 1	E L			_	e			F 0	T D		H D	EVE	57
							CCCCCEC		CCARC	മരവത	CORC	2002	ACCCATA	ים שרי		GAGGTAGA	
CIGIC	370	INGAN	380	CAAGA	390	-AGCAC	400	MCA	410	CCGII	420	JAGCAA	430	C11C	440	450	
	3,0		300		390		400		410		420		430		440	450	
T K	E A			P (		D 10	V D			~ ¥		м ъ			ь т	DAS	87
																GATGCCAGO	•
ACCAA.	460	CCGAG	470	GAAIC	480	JAIAC	490	.GCCC	500	CAAAA	510	110000	520	cruc.	530	540	
	400		4/0		400		450		300		3.0		320		330	310	
w D	E ·	T V	T A	6 1	ם י	s 6	N G	v	n v	0 0	ח	E L	PV	Y (	o 6	ETT	117
																GAGACCACO	
	550	LIIAC	560	Junn	570		580	.0100	590	۵	600		610		620	630	
	330		300		3.0		300		330		•••		• • •				
o s	L (	Q D	Y L	м 1	1 0	V E	T. T	P	FT	р т	מי	R A	I A	T :	s I	V D A	147
CAAAG	CORCO	ACCAT!		ATCTO	ccyc	TTTGA	ACTTACC	CCAT	TCACC	CATAC	CCATC	GCGCC	ATCGCG			GTCGATGC	•
	640		650		660		670		680		690		700		710	720	)
	040		050		000		0.0		000		0,0						
V D	ъ,	r c	Y L	T 1	rs	V E	D I	v	F S	T G	ם	D E	T G	L :	E E	VEA	177
GTTGA'	TGATA	CCGGC	TACCTG	ACGAT	TTTCT(	STCGA	AGACATO	GTGG	AAAGT	ATTGG	CGAC	SATGAA	ATCGGA	CTTG	AAGAA	GTTGAAGC	;
	730		740		750		760		770		780		790		800	810	)
V L	K	R T	O R	FI	) P	V G	V A	A	K D	L R	D	C L	L V	Q :	L S	QFA	207
GTTCT	CAAGC	GCATT											CTGGTT	CAGC	TTTCA	CAGTTTGC	:
	820		830		840		850		860		870		880		890	900	)
					• • • •												
K E	T	P W	I E	E /	A R	L I	I S	D	H L	D L	L	A N	H D	F !	R S	LMR	237
AAAGA	GACGC	CGTGG.	ATTGAA	GAAG	CCGC	CTGAT	CATCAGO	GATC	ATCTC	GATCT	GCTG	CCAAC	CACGAC	TTCC	GCAGC	CTGATGCG	:
	910		920		930		940		950		960		970		980	990	)
V T	R	LK	E E	V I	L K	E A	V N	L	1 0	S L	D	PR	P G	Q :	S I	Q T G	267
CTAAC	ccccc	TCAAA		CECE		CARCCO				macam		200000		~ ~ ~			•
			UMMUM		LAAAG		GGTAAAT	CIGA	TCCAA'	10001	GGAT		CCGGGGA	CAGT	CGATC	CAAACTGG	
GIAAC	1000	CAAA	1010	GIGI.	1020	JAAGC	1030	CIGA	TCCAA' 1040	ICGCI	1050		1060	CAGT	CGATC 1070	CAAACTGG( 1080	í
GIAAC		CAAA			1020		1030		1040		1050		1060	CAGT	1070	1080	)
	1000				1020				1040		1050		1060		CGATO 1070 S L	1080 P R L	í 297
E P	1000 E	y V	1010 I P	D 1	1020 , L	V R	1030 K V	N	1040 D R	w v	1050 V	E L	1060 N S AATTCA	D GATA	1070 S L GCCT1	1080 PRL PCCGCGCCTO	297
E P	1000 E	y V	1010 I P	D 1	1020 , L	V R	1030 K V	N CAACG	1040 D R	w v TGGGT	1050 V	E L	1060 N S	D GATA	1070 S L	1080 P R L	297
E P	1000 E AGAAT	y V	1010 I P ATTCCT	D V	1020 / L TTCTG0 1110	V R GTGCG	1030 K V TAAAGTO 1120	N CAACG	1040 D R ATCGT 1130	w v TGGGT	1050 V GGTT0 1140	E L GAGCTC	1060 N S AATTCA 1150	D GATA	1070 S L GCCT1 1160	1080 PRL PCCGCGCCTO	297 ;
E P GAGCC	1000 E AGAAT 1090	Y V ATGTC	I PATTCCT	D V	1020 / L PTCTG0 1110	V R GTGCG:	1030 K V TAAAGTC 1120 T R	N CAACG N	1040 D R ATCGT 1130	w v TGGGT	1050 V GGTT 1140	E L GAGCTC	1060 N S AATTCA 1150 R S	D GATA	1070 S L GCCT1 1160 L Q	PRL PCCGCGCCTC 1170 EAR	297 3 3 327
E P GAGCC	1000 E AGAAT 1090	Y V ATGTC	I PATTCCT	D V	1020 / L PTCTG0 1110	V R GTGCG:	1030 K V TAAAGTC 1120 T R	N CAACG N	1040 D R ATCGT 1130	W V TGGGT D G GACGG	1050 GGTTC 1140 QCCAG	E L GAGCTC	N S AATTCA 1150 R S CGTAGC	D GATA N AACC	1070 S L GCCTI 1160 L Q TGCAG	PRL PCCGCGCCCTC 1170 EAR GGAAGCGCGC	297 3 3 327
E P GAGCC	1000 E AGAAT 1090	Y V ATGTC. Q Q AGCAG	I PATTCCT	D V	1020 / L PTCTG0 1110	V R GTGCG:	1030 K V TAAAGTC 1120 T R	N CAACG N CAATG	1040 D R ATCGT 1130	W V TGGGT D G GACGG	1050 V GGTT 1140	E L GAGCTC	1060 N S AATTCA 1150 R S	D GATA N AACC	1070 S L GCCT1 1160 L Q	1080 PRL PCCGCGCCTO	297 3 3 327
E P GAGCC	1000 E S AGAAT 1090 N C	Y V ATGTC. Q Q AGCAG	I PATTCCT 1100 Y ATTGCC 1190	D NGACGT	1020 / L PTCTGG 1110 / G PGGGTA 1200	V R GTGCG N S AACAG	1030 K V FAAAGTO 1120 T R CACGCGO 1210	N CAACG N CAATG	D R ATCGT 1130 D A ACGCT 1220	W V TGGGT D G GACGG	1050 V GGTTC 1140 Q GCCAG 1230	E L GAGCTC F I TTTATC	N S AATTCA 1150 R S CGTAGC 1240	D GATA N AACC	1070 S L GCCTT 1160 L Q TGCAG	PRL PCCGCGCCTC 1170 EAR GGAAGCGCGC	297 3 327
E P GAGCCA K I AAGAT	AGAATA 1090 N CAATC 1180	Y V ATGTC. Q Q AGCAG	I PATTCCT 1100 Y ATATGCC 1190	D V CACCI A ! CCTAC	1020  L TTCTGG 1110  G TGGGTI 1200	V R GTGCG N S AACAGO	T R CACGCGG 1210  L L	N CAACG N CAATG	1040 D R ATCGT 1130 D A ACGCT 1220	W V TGGGT D G GACGG	V GGTTC 1140 CCAG 1230	E L GAGCTO F I FTTATO	N S AATTCA 1150 R S CGTAGC 1240	D GATA N AACC	1070 S L GCCTT 1160 L Q TGCAG 1250	PRL PCCGCGCCTC 1170 EAR GGAAGCGCGC 1260 FEQ	297 327 327
E P GAGCCA K I AAGAT	AGAATA 1090 N CAATC 1180	Y V ATGTC. Q Q AGCAG	I PATTCCT 1100 Y ATATGCC 1190	D V CACCI A ! CCTAC	1020  L TTCTGG 1110  G TGGGTI 1200	V R GTGCG N S AACAGO	K V TANAGTO 1120  T R CACGCGC 1210  L L	N CAACG N CAATG R GCGCG	D R ATCGT 1130 D A ACGCT 1220 V S TCAGC	W V TGGGT  D G GACGG  R C CGCTG	V GGTTC 1140 CCAG 1230	E L GAGCTO F I FTTATO	N S AATTCA 1150  R S CGTAGC 1240  Q Q CAGCAG	D GATA N AACC	S L GCCTT 1160 L Q TGCAG 1250 A F	PRL CCGCGCCCC 1177 EAR GGAAGCGCG 1260 FEQCTTTGAACA	297 327 357
E P GAGCCA K I AAGAT	AGAATA 1090 N CAATC 1180	Y V ATGTC. Q Q AGCAG	I PATTCCT 1100 Y ATATGCC 1190	D V CACCI A ! CCTAC	1020  L TTCTGG 1110  G TGGGTI 1200	V R GTGCG N S AACAGO	T R CACGCGG 1210  L L	N CAACG N CAATG R GCGCG	1040 D R ATCGT 1130 D A ACGCT 1220	W V TGGGT  D G GACGG  R C CGCTG	V GGTTC 1140 CCAG 1230	E L GAGCTO F I FTTATO	N S AATTCA 1150 R S CGTAGC 1240	D GATA N AACC	1070 S L GCCTT 1160 L Q TGCAG 1250	PRL PCCGCGCCTC 1170 EAR GGAAGCGCGC	297 327 357
E P GAGCC. K I AAGAT W L TGGCT	1000 E AGAAT 1090 N CAATC 1180 I GATCA 1270	Y V ATGTC. Q Q AGCAG K S AGAGC	I PATTCCT 1100 Y A TATGCC 1190 L E CTGGAG 1280	D Negacon	1020 / L PTCTGG 1110 / G PGGGTZ 1200 R N GCAACC 1290	V R GTGCG  N S AACAG  D T GACAC	T RCACGCGC 1210  L LCCTGCTG 1300	N N CAATG R GCGCG	D R ATCGT 1130 D A ACGCT 1220 V S TCAGC	W V TGGGT D G GACGG R C CGCTG	OCCAGE 1230 LITATOR TATOR	E L GAGCTC F I TTTATC V E GTCGAA	N S AATTCA 1150 R S CGTAGC 1240 Q Q CAGCAG	D GATA N AACC Q CAGG	1070 S L GCCTT 1160 L Q TGCA6 1250 A F CTTTT 1340	PRL CCGCGCCTC 1170 EAR GGAAGCGCG 1260 FEQ TTTTGAACA 1350	297 327 357
E PGAGCC	1000 E AGAAT: 1090 N CAATC: 1180 I GATCA: 1270	Y V ATGTC. Q Q AGCAG K S AGAGC	1010 I P ATTCCT 1100 Y A TATGCC 1190 L E CTGGAG 1280 K P	D N GACG	1020  / L PTCTGG 1110  / G PGGGTI 1200  R N GCAACG 1290  / L	V R GTGCG	T RCACGCGC	N N CAATG R GCGCG	D R ATCGT 1130 D A ACGCT 1220 V S TCAGC 1310	W VTGGGT D GGACGG	1050 V GGTTC 1140 CCAG 1230 I I	E L GAGCTC F I TTTATC V E GTCGAA	N S AATTCA 1150  R S CGTAGC 1240  Q Q CAGCAG 1330  T I	D GATA N AACC Q CAGG	S L GCCTT1160 L Q TGCAG1250 A F CTTTT1340 R V	PRL PCCGCGCCTC 1170 EAR GGAAGCGCGC 1260 FEQ PTTTGAACA 1350 TTQ	297 327 357 357
E PGAGCC	1000 E AGAATC 1090 N CAATC 1180 I GATCA 1270 E AGAGT	Y V ATGTC. Q Q AGCAG K S AGAGC	I PATTCCT 1100 Y A TATGCC 1190 L E CTGGAG 1280 K PAAAACCC	D N GACG	1020  L TTCTG( 1110  G TGGGTI 1200  R N GCAAC( 1290  L TGCTGG	V R GTGCG	K V TAAAGTC 1120 T R CACGCGC 1210 L L CCTGCTG 1300 I A TATACGCT	N N CAATG R GCGCG	1040 D R ATCGT 1130 D A ACGCT 1220 V S TCAGC 1310 A V CCGTC	W VTGGGT D GGACGG R CCGCTG	OCCAGE TATCE	E L GAGCTC F I TTTATC V E GTCGAA	N S AATTCA 1150 R S CGTAGC 1240 Q Q CAGCAG 1330 T I ACTATT	D GATA N AACC Q CAGG	S L GCCTT 1160 L Q TGCAG 1250 A F CTTTT 1340 R V GCGTT	PRL CCGCGCCTT 1170 EAR GGAAGCGCGG 1260 FEQ FTTTGAACA TTQ CACTACGCA	297 327 357 357
E PGAGCC	1000 E AGAAT: 1090 N CAATC: 1180 I GATCA: 1270	Y V ATGTC. Q Q AGCAG K S AGAGC	1010 I P ATTCCT 1100 Y A TATGCC 1190 L E CTGGAG 1280 K P	D N GACG	1020  / L PTCTGG 1110  / G PGGGTI 1200  R N GCAACG 1290  / L	V R GTGCG	T RCACGCGC	N N CAATG R GCGCG	D R ATCGT 1130 D A ACGCT 1220 V S TCAGC 1310	W VTGGGT D GGACGG R CCGCTG	1050 V GGTTC 1140 CCAG 1230 I I	E L GAGCTC F I TTTATC V E GTCGAA	N S AATTCA 1150  R S CGTAGC 1240  Q Q CAGCAG 1330  T I	D GATA N AACC Q CAGG	S L GCCTT 1160 L Q TGCAG 1250 A F CTTTT 1340 R V	PRL PCCGCGCCTC 1170 EAR GGAAGCGCGC 1260 FEQ PTTTGAACA 1350 TTQ	297 327 357 357
E P GAGCC K I AAGAT W L TGGCT G E GGTGA	1000 E AGAAT 1090 N CAATC 1180 I GATCA 1270 E AGAGT	Y V ATGTC. Q Q AGCAG K S AGAGC F M TTATG	I PATTCCT 1100 Y A TATGCC 1190 L E CTGGAG 1280 K PAAAACCC 1370	D N GACGT A II GCTAT S II GAGCCO M N GATGGT	1020 / L PTCTGG 1110 / G PGGGTI 1200 R N GCAACG 1290 / L TGCTGG	V R GTGCG  N S AACAGO  D T GACACO	K V TAAAGTC 1120 T R CACGCGC 1210 L L CCTGCTG 1300 I A TTATCGCT 1390	N PAACG N PAATG R GCGCG Q PCAGG	D R ATCGT 1130 D A ACGCT 1220 V S TCAGC 1310 A V CCGTC	W V TGGGT  D G GACGG  R C CGCTG	GGTTC 1140 CCAG 1230 ETATC 1320 HGCATC	E L GAGCTC F I TTTATC V E GTCGAA E S GAATCC	N S AATTCA 1150 R S CGTAGC 1240 Q Q CAGCAG 1330 T I ACTATT	D GATA N AACC CAGG CAGG	1070 S L GCCTT 1160 L Q TGCAG 1250 A F CTTTT 1340 R V GCGTT 1430	PRL CCGCGCCTC 1170 EAR GGAAGCGCGC 1260 FEQ TTTTGAACA 1350 TTQ TACTACGCA 1440	297 327 357 357
E P GAGCC.  K I AAGAT  W L TGGCT  G E GGTGA	AGAATCA 1270 E AGAGTCA 1270 E AGAGTCA 1360 L	Y V ATGTC Q Q AGCAG K S AGAGC F M TTATG	I PATTCCT 1100 Y A TATGCC 1190 L E CTGGAG 1280 K PAAAACCC 1370 P R	D NO GACGO	1020  / L PTCTGG 1110  4 G PGGGTI 1200  R N GCAACC 1290  / L PGCTGG 1380  I F	V R GTGCG	X V TAAAGTC 1120  T R CACGCGC 1210  L L CCTGCTG 1300  I A TATCGCT 1390  K Y	N CAACG N CAATG CCGCG CCGCG	1040 D R ATCGT 1130 D A ACGCT 1220 V S TCAGC 1310 A V CCGTC 1400 F S	W V TGGGT D G GACGG R C CGCTG E M GAAAT	TOSO VGGTTC 1140 CCCAG 1230 CITATCC 1320 CITATCC 1320 CITATCC 1410	E L GAGCTC F I TTTATC V E GTCGAA E S GAATCC	N S AATTCA 1150  R S CGTAGC 1240  Q Q CAGCAG 1330  T I ACTATT 1420  E G	D GATA N AACC CAGG TCAC	1070 S L GCCTT 1160 L Q TGCAG 1250 A F CTTTT 1340 R V GCGTT 1430 G E	PRL TCCGCGCGCCTT 1170  EAR GGAAGCGCG 1260  FEQ TTTTGAACA 1350  TTQ CACTACGCA 1440  ASS	297 327 357 357 387
E P GAGCC.  K I AAGAT  W L TGGCT  G E GGTGA	AGAATCA 1180  I GATCA 1270  E AGAGT 1360  L CCTGC	Y V ATGTC Q Q AGCAG K S AGAGC F M TTATG	1010 I PATTCCT 1100 Y A TATGCC 1190 L E CTGGAG 1280 K P AAACCC 1370 P R CCACGC	D NO GACGO	1020  / L PTCTGG 1110  4 G PGGGTI 1200  R N GCAACC 1290  / L PGCTGG 1380  I F PTTTTC	V R GTGCG	K V TAAAGTC 1120  T R CACGGGC 1210  L L CCTGCTG 1300  I A TATCGCT 1390  K Y GAAGTAT	N CAATG R GCGCG TCAGG	D R ATCGT 1130 D A ACGCT 1220 V S TCAGC 1310 A V CCGTC 1400 F S TCTCC	W V TGGGT D G GACGG R C CGCTG E M GAAAT S H AGCCA	1050 V GGTTC 1140 CCCAG 1230 CIATO 1320 I H GCATC 1410 I V	E L GAGCTC F I TTTATC V E GTCGAA E S GAATCC	N S AATTCA 1150 R S CGTAGC 1240 Q Q CAGCAG 1330 T I ACTATT 1420 E G GAAGGC	D GATA N AACC CAGG TCAC	S L GCCTT 1160 L Q TGCAG 1250 A F CTTTT 1340 R V GCGTT 1430 G E GCGA	PRL PCCGCGCGCT 1177 EAR GGAAGCGCG 1260 FEQ FTTTGAACA 1350 TTQ FACTACGCA 1440 ASSACCATCGTCC	297 327 357 357 387
E P GAGCC.  K I AAGAT  W L TGGCT  G E GGTGA	AGAATCA 1270 E AGAGTCA 1270 E AGAGTCA 1360 L	Y V ATGTC Q Q AGCAG K S AGAGC F M TTATG	I PATTCCT 1100 Y A TATGCC 1190 L E CTGGAG 1280 K PAAAACCC 1370 P R	D NO GACGO	1020  / L PTCTGG 1110  4 G PGGGTI 1200  R N GCAACC 1290  / L PGCTGG 1380  I F	V R GTGCG	X V TAAAGTC 1120  T R CACGCGC 1210  L L CCTGCTG 1300  I A TATCGCT 1390  K Y	N CAATG R GCGCG TCAGG	1040 D R ATCGT 1130 D A ACGCT 1220 V S TCAGC 1310 A V CCGTC 1400 F S	W V TGGGT D G GACGG R C CGCTG E M GAAAT S H AGCCA	TOSO VGGTTC 1140 CCCAG 1230 CITATCC 1320 CITATCC 1320 CITATCC 1410	E L GAGCTC F I TTTATC V E GTCGAA E S GAATCC	N S AATTCA 1150  R S CGTAGC 1240  Q Q CAGCAG 1330  T I ACTATT 1420  E G	D GATA N AACC CAGG TCAC	1070 S L GCCTT 1160 L Q TGCAG 1250 A F CTTTT 1340 R V GCGTT 1430 G E	PRL TCCGCGCGCCTT 1170  EAR GGAAGCGCG 1260  FEQ TTTTGAACA 1350  TTQ CACTACGCA 1440  ASS	297 327 357 357 387
E PGAGCC	1000 E AGAATC 1090 N CAATC 1180 I GATCA 1270 E AGAGT 1360 L CCTGC 1450	Y V ATGTC. Q Q AGCAG K S AGAGC F M TTATG H S ACAGT	1010 I PATTCCT 1100 Y A TATGCC 1190 L E CTGGAG 1280 K P AAACCC 1370 P R CCACGG 1460	D VICE CONTROL OF CONT	1020  / L  TTCTGG 1110  4 G  PGGGT/ 1200  R N  GCAACC 1290  / L  PGCTGC 1380  I F  TTTTTC 1470	V R GTGCG  N S AACAG  D T GACAC  A D GCGGA	K V TAAAGTC 1120  T R CACGCGC 1210  L L CCTGCTG 1300  I A TATCGCT 1390  K Y GAAGTAT 1480	N CAATG R GCGCG CCAGG	D R ATCGT 1130 D A ACGCT 1220 V S TCAGC 1310 A V CCGTC 1400 F S TCTCC	W V TGGGT  D G GACGG  R C CGCTG  E M GAAAT  S H AGCCA	TOSO VGGTTC 1140 CCAG 1230 CTATCC 1320 HGCATC 1410 VGTGTG 1500	E L GAGCTC F I TTTATC V E GTCGAA E S GAATCC	N S AATTCA 1150 R S CGTAGC 1240 Q Q CAGCAG 1330 T I ACTATT 1420 E G GAAGGC 1510	D GATA N AACC CAGG CAGG	1070 S L GCCTT 1160 L Q TGCAG 1250 A F CTTTT 1340 R V GCGTT 1430 G E GCGA	PRL PCCGCGCCTC 1177 EAR GGAAGCGCG 1260 1260 TT T Q PACTACGCAA 1440 AS S AGCATCGTCC 1530	297 327 357 357 387
E PGAGCC	AGAATCA 1090  N CAATCA 1180  I GATCA 1270  E AGAGT 1360  L CCTGC 1450	Y V ATGTC. Q Q AGCAG K S AGAGC F M TTATG H S ACAGT	1010 I PATTCCT 1100 Y A TATGCC 1190 L E CTGGAG 1280 K P AAAACCC 1370 P R CCACGC 1460 L V	D N GACGTAT S I GACCC M N GATGGT GCTAT	1020 / L PTCTGG 1110 / G PGGGTI 1200 R N GCAACC 1290 / L PGCTGC 1380 I F PTTTTC 1470 K L	V R GTGCGT  N S AACAGO  D T GACACO  A D GCGGAT  E L GAGCTO	K V TANAGTC 1120 T R CACGCGC 1210 L L CCTGCTG 1300 I A TATCGCT 1390 K Y GAAGTAT 1480 A E	N CAACG N CAATG R CCGCG CCAGG	1040 D R ATCGT' 1130 D A ACGCT' 1220 V S TCAGC' 1310 A V CCGTC' 1400 F S TCTCC.	W V TGGGT  D G GACGG  R C CGCTG  E M GAAAT  S H AGCCA	TOSO  GGTTC 1140  GCCAG 1230  TATCC 1320  HCCATC 1410  VTGTG 1500	E L GAGCTC F I FTTATC V E GTCGAA E S GAATCC	N S AATTCA 1150  R S CGTAGC 1240  Q Q CAGCAG 1330  T I ACTATT 1420  E G GAAGGC 1510  S K	D GATA	1070 S L GCCTT 1160 L Q TGCAG 1250 A F CTTTT 1340 R V GCGTT 1430 G E GCGAA 1520 T T	PRL rccgcgccrr 1177 EAR GGAAGCGCGC 1266 FEQ TTTTGAACAJ 1356 TTQ TACTACGCAL 1440 ASS AGCATCGTCC 1530 MLS	297 327 357 357 387 417
E PGAGCC	AGAATCA 1090  CAATCA 1180  I GATCA 1270  E AGAGT 1360  L CCTGC 1450  I CATTC	Y V ATGTC. Q Q AGCAG K S AGAGC F M TTATG H S ACAGT	1010 I PATTCCT 1100 Y A TATGCC 1190 L E CTGGAG 1280 K P AAAACCC 1370 P R CCACGC 1460 L V CTGGTC	D N GACGTAT S I GACCC M N GATGGT GCTAT	1020  L TTCTGC 1110  G GGGT/ 1200  R N GCAACC 1290  V L TGCTGC 1380  I F TTTTTC 1470  K L AATTA/	V R GTGCGT  N S AACAGO  D T GACACO  A D GCGGAT  E L GAGCTO	K V TANAGTO 1120  T R CACGCGC 1210  L L CCTGCTG 1300  I A TATCGCT 1390  K Y GAAGTAT 1480  A E CCCCGAA	N N CAATG R R CCCCG Q CTCAGG F F TTTTCT N N NAAACC	D R ATCGTT 1130  D A ACGCTT 1220  V S TCAGC 1310  A V CCGTC 1400  F S TCTCC 1490  P A CCGCG	W V TGGGT  D G GACGG  R C CGCTG  E M GAAAT  S H AGCCA	OF THE PROPERTY OF THE PROPERT	E L GAGCTC F I FTTATC V E GTCGAA E S GAATCC	N S AATTCA 1150 R S CGTAGC 1240 Q Q CAGCAG 1330 T I ACTATT 1420 E G GAAGGC 1510 S K AGTAAG	D GATA  N AACC CAGG STCAC GGGCG	S L GCCTT 1160 L Q TGCAG 1250 A F CTTTT 1340 R V GCGTT 1430 G E GCGAA 1520 T T CCACC	PRL rccgccgccrr 1177  EAR GGAAGCGCG 1260  FEQ rTTTGAACA 1350  TTQ rACTACGCA 1440  ASS AGCATCGTCC 1530  MLS CATGCTATCC	297 327 357 357 387 417
E PGAGCC	AGAATCA 1090  N CAATCA 1180  I GATCA 1270  E AGAGT 1360  L CCTGC 1450	Y V ATGTC. Q Q AGCAG K S AGAGC F M TTATG H S ACAGT	1010 I PATTCCT 1100 Y A TATGCC 1190 L E CTGGAG 1280 K P AAAACCC 1370 P R CCACGC 1460 L V	D N GACGTAT S I GACCC M N GATGGT GCTAT	1020 / L PTCTGG 1110 / G PGGGTI 1200 R N GCAACC 1290 / L PGCTGC 1380 I F PTTTTC 1470 K L	V R GTGCGT  N S AACAGO  D T GACACO  A D GCGGAT  E L GAGCTO	K V TANAGTC 1120 T R CACGCGC 1210 L L CCTGCTG 1300 I A TATCGCT 1390 K Y GAAGTAT 1480 A E	N N CAATG R R CCCCG Q CTCAGG F F TTTTCT N N NAAACC	1040 D R ATCGT' 1130 D A ACGCT' 1220 V S TCAGC' 1310 A V CCGTC' 1400 F S TCTCC.	W V TGGGT  D G GACGG  R C CGCTG  E M GAAAT  S H AGCCA	TOSO  GGTTC 1140  GCCAG 1230  TATCC 1320  HCCATC 1410  VTGTG 1500	E L GAGCTC F I FTTATC V E GTCGAA E S GAATCC	N S AATTCA 1150  R S CGTAGC 1240  Q Q CAGCAG 1330  T I ACTATT 1420  E G GAAGGC 1510  S K	D GATA  N AACC CAGG STCAC GGGCG	1070 S L GCCTT 1160 L Q TGCAG 1250 A F CTTTT 1340 R V GCGTT 1430 G E GCGAA 1520 T T	PRL rccgcgccrr 1177 EAR GGAAGCGCGC 1266 FEQ TTTTGAACAJ 1356 TTQ TACTACGCAL 1440 ASS AGCATCGTCC 1530 MLS	297 327 357 357 387 417
E P GAGCC.  K I AAGAT.  W L TGGCT.  G E GGTGA.  K Y AAATA.  T A ACGGC.	1000 E AGAAT: 1090 N CAATC. 1180 1270 E AGAGT: 1360 L CCTGC. 1450	Y V ATGTC. Q Q AGCAG K S AGAGC F M TTATG H S ACAGT R A GCGCG	I PATTCCT 1100 Y A TATGCC 1190 L E CTGGAG 1280 K P AAACCC 1370 P R CCACGC 1460 L V CTGGTG 1550	D V GGACGTA  A 1 S 1 SAGCCO  M 1 SATGGTA  K 1 SAAGA	1020  L PTCTGC 1110  G GGGTL 1200  R N GCAACC 1290  L PGCTGC 1380  L PGTTTTC 1470  K L AATTAL 1560	V R GTGCG  N S AACAG  D T GACAC  A D GCGGA  E L GAGCT  I A ATCGCC	K V TRANAGTC 1120 T R CACGCGC 1210 L L CCTGCTG 1300 I A TATCGCT 1390 K Y GAAGTAT 1480 A E CGCGGAA 1570	N NAACG N R R R CCGCGCG Q Q Q CCAGGG F F TTTCT	1040 D R ATCGT 1130 D A ACGCT 1220 V S CTCAGC 1400 F S TCTCC 1490 P A CCGCG 1580	W V TGGGT D G GGACGG R C CGCTG E M GAAAAT S H AGCCA K P AAAACC	OCCAGGINATION OF THE PROPERTY	E L GAGCTC F I TTTATC V E GTCGAA E S GAATCC N T AACACC	N S AATTCA 1150 R S CGTAGC 1240 Q Q CAGCAG 1330 T I ACTATT 1420 E G GAAGGC 1510 S K AGTAAG	D GATA N AACC CAGG S TCAC G GGCG	1070 S L GCCTT 1160 L Q TTGCAG 1250 A F CTTTT 1340 R V GCGTT 1430 G E GCGAA 1520 T T CCCACC 1610	PRL rccgccgccrr 1177 EAR GGAAGCGGG 1266 FEQ PTTTGAACA1 1351 TTQ PACTACGCA1 1440 ASS AGCATCGTCC 1530 MLS CATGCTATCC 1620	297 327 357 357 387 417
E P GAGCC.  K I AAGAT.  W L TGGCT.  G E GGTGA.  K Y AAATA.	1000 E AGAAT 1090 N CAATC 1180 1270 E AGAGT 1360 L CCTGC 1450 1	Y V ATGTC. Q Q AGCAG K S AGAGGC F M TTATG H S ACAGT	I PATTCCT 1100 Y A TATGCC 1190 L ECTGGAG 1280 K PAAACCC 1370 P R CCACGC 1460 L V CTGGTC 1550	D V GACGO A 11 GCCTA S 1 GAGCCC M V GATGGC G : GAGGTA K 1 GAAGA R 1	1020 / L PTCTGG 1110 / G PGGGTI 1200 R GCAACC 1290 / L PGCTGG 1380 / L PTTTTC 1470 K L AATTA	V R GTGCGC N S AACAGC D T GACACC A D GCGGA E L GAGCTC I A ATCGCC	T RCACGCGC 1210  L L CCTGCTG 1300  I A TATCGCT 1390  V GAAGTAT 1480  A E CGCGGAA 1570  K Y	N PAACG  R R GCGCGG  CAGGG  P F TTTCT  N NAACC	1040 D R ATCGT 1130 D A ACGCT 1220 TCAGC 1310 A V CCGTC 1400 F S TCTCC 1490 P A CCGCG 1580 E S	W V TGGGT  D G GACGG  R C C GCCTG  E M GAAAT  S H AAACCA	OCCAGGOTTON 1140 CCCAGGOTTON 1230 CCCAGGOTTON 1230 CCCAGGOTTON 1410 CCCAGG	E L GAGCTC F I FTTATC V E GTCGAA E S GAATCC N T AACACC	1060  N S AATTCA 1150  R S CGTAGC 1240  Q Q CAGCAG 1330  T I ACTATT 1420  E G GAAGGC 1510  S K AGGTAAAG 1600  S N	D GATA	1070 S L GCCTT 1160 L Q TTGCAG 1250 A F CTTTT 1340 R V GGCGTT 1430 T T CCACCC 1610 R K	PRL PCCGCGCCTCTCTTTTTTTTTTTTTTTTTTTTTTTTTTT	327 327 357 357 387 417 447
E P GAGCC.  K I AAGAT.  W L TGGCT.  G E GGTGA.  K Y AAATA.	1000 E AGAAT. 1090 N CAATC. 1180 1270 E AGAGT 1360 L CCTGC. 1450 I GGGGTA GGGGGTA	Y V ATGTC.  Q Q AGCAG  K S AGAGC  F M TTTATG  H S ACAGT  R A GCGCG	1010 I PATTCCT 1100 Y A TATGCC 1190 L E CTGGAC 1280 K P AAAACC 1370 P R CCACGC 1460 L V CTGGTC 1550 V A GTGGCA	D V GACGO A 11 GCCTA S 1 GAGCCC M V GATGGC G : GAGGTA K 1 GAAGA R 1	1020  L PTCTGG 1110  G GAAC 1290  L PCCTGG 1380  L PCTTTTT 1470  K L AATTAA	V R GTGCGC N S AACAGC D T GACACC A D GCGGA E L GAGCTC I A ATCGCC	T R CACGCGC 1210  L L CCTGCTG 1300  I A TATCGCT 1390  K Y GAAGTAT 1480  A E CGCGGAA 1570  K Y TAAGTAC	N N CAATG N R R GCGCG C C C C C C C C C C C C C C C C C	1040 D R ATCGT 1130 D A ACCGCT 1220 1220 1400 F S TCTCC 1490 P A C1580	W V TGGGT D G GACGG R C CGCTG E M GGAAAT S H AGCCA K P L S TTTATC	OCCAGO THE CONTROL OF	E L GAGCTC F I FTTATC V E GTCGAA E S GAATCC N T AACACC	N S AATTCA N S AATTCA 1150 R S CGTAGC 1240 Q Q CAGCAG 1330 T I ACTATT 1420 E G GAAGGC 1510 S K AGTAAG 1600 S N TCAAAAT	D GATA  N N AACC  CAGG  CAGG  CCAGG  CCTCA  Q CCAGC	1070 S L GCCTT 1160 L Q TTGCAG 1250 A F CTTTT 1340 R V TGCAG 1520 T T CCACC 1610 R K GCAAA	PRL rccgccgccrr 1177 EAR GGAAGCGGG 1266 FEQ PTTTGAACA1 1351 TTQ PACTACGCA1 1440 ASS AGCATCGTCC 1530 MLS CATGCTATCC 1620	327 327 357 387 387 417 447
E P GAGCC.  K I AAGAT.  W L TGGCT.  G E GGTGA.  K Y AAATA.	1000 E AGAAT 1090 N CAATC 1180 1270 E AGAGT 1360 L CCTGC 1450 1	Y V ATGTC.  Q Q AGCAG  K S AGAGC  F M TTTATG  H S ACAGT  R A GCGCG	I PATTCCT 1100 Y A TATGCC 1190 L ECTGGAG 1280 K PAAACCC 1370 P R CCACGC 1460 L V CTGGTC 1550	D V GACGO A 11 GCCTA S 1 GAGCCC M V GATGGC G : GAGGTA K 1 GAAGA R 1	1020 / L PTCTGG 1110 / G PGGGTI 1200 R GCAACC 1290 / L PGCTGG 1380 / L PTTTTC 1470 K L AATTA	V R GTGCGC N S AACAGC D T GACACC A D GCGGA E L GAGCTC I A ATCGCC	T RCACGCGC 1210  L L CCTGCTG 1300  I A TATCGCT 1390  V GAAGTAT 1480  A E CGCGGAA 1570  K Y	N N CAATG N R R GCGCG C C C C C C C C C C C C C C C C C	1040 D R ATCGT 1130 D A ACGCT 1220 TCAGC 1310 A V CCGTC 1400 F S TCTCC 1490 P A CCGCG 1580 E S	W V TGGGT D G GACGG R C CGCTG E M GGAAAT S H AGCCA K P L S TTTATC	OCCAGGOTTON 1140 CCCAGGOTTON 1230 CCCAGGOTTON 1230 CCCAGGOTTON 1410 CCCAGG	E L GAGCTC F I FTTATC V E GTCGAA E S GAATCC N T AACACC	1060  N S AATTCA 1150  R S CGTAGC 1240  Q Q CAGCAG 1330  T I ACTATT 1420  E G GAAGGC 1510  S K AGGTAAAG 1600  S N	D GATA  N N AACC  CAGG  CAGG  CCAGG  CCTCA  Q CCAGC	1070 S L GCCTT 1160 L Q TTGCAG 1250 A F CTTTT 1340 R V GGCGTT 1430 T T CCACCC 1610 R K	PRL PCCGCGCGCTT 1177  EAR GGAAGCGCGC 1266  FEQ PTTTGAACAJ 1350  TTQ PACTACGCAL 1440  ASS AGCATCGTC 1530  MLS CATGCTATCC 1620  QLV GCAGCTGGT	327 327 357 387 387 417 447
E P GAGCC.  K I AAGAT.  W L TGGCT.  G E GGTGA.  K Y AAATA.	1000 E AGAAT. 1090 N CAATC. 1180 1270 E AGAGT 1360 L CCTGC. 1450 I GGGGTA GGGGGTA	Y V ATGTC.  Q Q AGCAG  K S AGAGC  F M TTTATG  H S ACAGT  R A GCGCG	1010 I PATTCCT 1100 Y A TATGCC 1190 L E CTGGAC 1280 K P AAAACC 1370 P R CCACGC 1460 L V CTGGTC 1550 V A GTGGCA	D V GACGO A 11 GCCTA S 1 GAGCCC M V GATGGC G : GAGGTA K 1 GAAGA R 1	1020  LTCTGGT 1110  G GGGAC 1290  V L 1380  I F FTTTT 1470  K L AATTAA 1560  R T GGAACC 1650	V R GTGGG  N S AACAGG  D T GACACG  A D GGGGA  E L GAGCT  I A ATCGC  V A	T RCACGCGC 1210  L L CCTGCTG 1300  I A TATCGCT 1390  K Y GAACTAT 1480  A E CGCGGAA 1570  K Y TAAGTAC 1660	N N CAACG N R CCGCG Q Q CCAGGG FTTTCT N LAACC	1040 D R ATCGT 11130 A N AGGGCT 11220 V S TCAGC 1310 A V F S TCTCC 1490 F S TCTCC	W V TGGGT  D G GACGG  R CC CCGCTG  E M GAAAT  S H AGCCA  K P AAACC	OCCAGO TATACO TA	E L GAGCTC F I TTTATC V E GTCGAA E S GAATCC N T AACACC	N S AATTCA 1150  R S CGTAGC 1240  Q Q CAGCAG 1330  T I ACTATT 1420  S K AGTAGC 1510  S K AGTAGC 1600  S N TCAAAT 1690	D GATA	1070 S L GCCTI 11160 L Q GCG1 1250 A F CTTTT 1340 R V GCGT1 1430 G E GCGAA 1520 T T T CCACC 1610 R K GCAAA 1700	PRL PCCGCGCGCT 1170  EAR GGAAGCGCGC 1260  FEQ FTTTGAACA 1350  TTQ FACTACGCA 1440  AAS AGCATCGTC 1530  MLS CATGCTATCC 1620  QLV GCAGCTGGT 1710	297 327 357 357 387 417 447
E P P GAGCC.  K I I AAAGAT.  W L TGGCT.  G E GGTGA.  T A AAATA.  T A ACGGC.	1000 E AGAAT. 1090 N CAATC. 1180 1270 E AGAGT 1360 I CATTC 1540 G G G G G G G G G G G G G G G G G G G	Y V ATGTC.  Q Q Q AGCAG K S AGAGAGC F M TTATG H S ACAGT R A GCGCG	I PATTCCT 11000 I PATTCCT 11000 YA A TATGCC 11190 L E CTGGACA 11280 K P AAACCC 1370 P R CCACGG 1460 L V CTGGTC 1550 V A AGTGGGC 1640	D V GACGO A I S I SAGCO G C GGTA' K I SAAGAA R I	1020  L TTCTGG 1110  G G C 1200  R N GCAACC 1290  L TCCTGG 1470  L	V R STGCG: N S S D T GAACAGG A D GCGGA: I A ATCGCG	X V TRANAGTC 1120 T R CACGCGC 1210 L L CCTGCTG 1300 I A TATCGCT 1390 K Y GAAGTAT 1480 A E CCGCGGAA 1570 K Y TAAGTAC	N CAACG  R R R CCGAGG  PCAGG  F TTTCT  N N N N N N N N N N N N N N N N	1040 D R ATCGT 11130 D A 1220 V S TCAGC 1310 A V CCGTC 1490 P A CCGCG 1589 E S 1670 H N	W V TGGGT D G GACGG R C CGCTG E M GAAAAT S H AGCCA K P AAAACC	V GGTTT 1140 C CCAGT 1230 I TATCC 1410 C CTAGT 1410 C CTAGT 1500 C L CTATTG 1500 C L CTATTG 1680 E I	E L JAGCTC  F I FTTATC  V E STCGAA  N T AACACC  S D AGTGAT	N S AATTCA 1150 R S CGTAGC 1240 Q Q CAGCAG 1330 T I ACTATT 1420 E G GAAGGC 1510 S K AGTAAG 1600 S N TCAAAAT 1690 A M	D GATA  N AACC  CAGG  S TCAC  GGGCG  L CTCA  CTCA  CAGC  R	1070 S L GCCTT 1160 L Q TTGCAG 1250 A F TTTT 1340 R V GCGTT 1430 G E GCGAA 1520 T T C1610 R K GCAAC	PRL rCCGCGCGCTT 1177  EAR GGAAGCGGG 1266  FEQ PTTTGAACA1 1351  TTQ PACTACGCA1 1446  ASS AGCATCGTCC 1536  MLS CATGCTATCC 1626  QLV GCAGCTGGTT 1710  FVT	297 327 357 357 357 387 417 447
E P P GAGCC.  K I I AAAGAT.  W L TGGCT.  G E GGTGA.  T A AAATA.  T A ACGGC.	1000 E AGAAT. 1090 I GARCA 1180 I GARCA 1270 CCTGC 1450 L CCTGC 1540 G GGGGTA 1630 CCAACC	Y V ATGTC.  Q Q Q AGCAG K S AGAGAGC F M TTATG H S ACAGT R A GCGCG	I PATTCCT I 1100 Y A TATGCC 1190 L E CTGGAACCC 1370 F P R P CCACGC 1460 L C CTGGAC 1640 V A GTGGCA 1640	D V GACGO A I S I SAGCO G C GGTA' K I SAAGAA R I	1020  L TTCTGG 1110  R GGAGC 1200  R CGAACC 1290  L CCTGG 1380  L FTTTTT 1470  R AAATTAA 1560  R T GAACC 1650	V R STGCG: N S S D T GAACAGG A D GCGGA: I A ATCGCG	K V TANAGTC 1120 T R CACGCGC 1210 L L CCTGCTG 1300 I A TATCGCT 1390 K Y GAAGTAT 1480 A E CGCGGAA 1570 K Y TAAGTAC 1660 N I T TAAGTAC	N CAACG N R R CCCAGG PTTTCT N AAACC	D R ATCGT 1130  D A ACGCT 1220  V S TCAGC 11400  F S TCTCC 1490  P CCGGG 1580  E S AGTCT H N CACAA	W V TGGGT D G GACGG R C CGCTG E M GAAAAT S H AGCCA K P AAAACC	V GGTTT 1140 C CCAGT 1230 I TATCC 1410 C CTAGT 1410 C CTAGT 1500 C L CTATTG 1500 C L CTATTG 1680 E I	E L JAGCTC  F I FTTATC  V E STCGAA  N T AACACC  S D AGTGAT	N S AATTCA 1150 R S CGTAGC 1240 Q Q CAGCAG 1330 T I ACTATT 1420 S K AGTAAG 1600 S N TCAAAAT 1690 M CGCGAAT	D GATA  N AACC  CAGG  S TCAC  G GGCG  CTCA  CTCA  G GGCG  R GCGCG	1070 S L GCCTT 1160 L Q TTGCAG 1250 A F TTTT 1340 R V GCGTT 1430 G E GCGAA 1520 T T C1610 R K GCAAC	PRL rccgcgccrt 1170 EAR GGAAGCGCGC 1260 FEQ FTTTGAACA 1350 TTQ FACTACGCA 1440 ASS AGCATCGTC 1530 MLS CATGCTATC 1620 QLV GCAGCTGGT 1710	297 327 357 357 367 417 447
E P P GAGCC.  K I I AAAGAT.  W L TGGCT.  G E GGTGA.  T A AAATA.  T A ACGGC.	1000 E AGAAT. 1090 N CAATC. 1180 1270 E AGAGT 1360 I CATTC 1540 G G G G G G G G G G G G G G G G G G G	Y V ATGTC.  Q Q Q AGCAG K S AGAGAGC F M TTATG H S ACAGT R A GCGCG	I PATTCCT 11000 I PATTCCT 11000 YA A TATGCC 11190 L E CTGGACA 11280 K P AAACCC 1370 P R CCACGG 1460 L V CTGGTC 1550 V A AGTGGGC 1640	D V GACGO A I S I SAGCO G C GGTA' K I SAAGAA R I	1020  L TTCTGG 1110  G G C 1200  R N GCAACC 1290  L TCCTGG 1470  L	V R STGCG: N S S D T GAACAGG A D GCGGA: I A ATCGCG	X V TRANAGTC 1120 T R CACGCGC 1210 L L CCTGCTG 1300 I A TATCGCT 1390 K Y GAAGTAT 1480 A E CCGCGGAA 1570 K Y TAAGTAC	N CAACG N R R CCCAGG PTTTCT N AAACC	1040 D R ATCGT 11130 D A 1220 V S TCAGC 1310 A V CCGTC 1490 P A CCGCG 1589 E S 1670 H N	W V TGGGT D G GACGG R C CGCTG E M GAAAAT S H AGCCA K P AAAACC	OF THE PROPERTY OF THE PROPERT	E L JAGCTC  F I FTTATC  V E STCGAA  N T AACACC  S D AGTGAT	N S AATTCA 1150 R S CGTAGC 1240 Q Q CAGCAG 1330 T I ACTATT 1420 E G GAAGGC 1510 S K AGTAAG 1600 S N TCAAAAT 1690 A M	D GATA  N AACC  CAGG  S TCAC  G GGCG  CTCA  CTCA  G GGCG  R GCGCG	1070 S L GCCTT11160 L Q GCTTTT11340 R V GCGTT11430 T T T CCACC1610 R K GCAAC11700 R C G C G G E G C G C G C G C G C G C G C	PRL PCCGCCCTCC 1177  EAR GGAAGCGCG 1260  FEQ FTTTGAACA 1350  TTQ FACTACGCA 1444  ASS AGCATCGTC 1534  MLS CATGCTATCC QLV GCAGCTGGT 1771  FVT FCGTTACCGTACCGT	297 327 357 357 367 417 447
E P GAGCC.  K I I AAGAT.  W L TGGCT.  G E GGTGA.  K Y AAATA.  T A ACGGC.  D Q GATCA.	1000 E ACACCC CAACCC T720	Y V ATGTC.  Q Q Q AGCAG K S AGAGC F M TTATG H S ACAGT R A GCGCG I M GGATAA	I PATTCCT I PATTCCT I PATTCCT I 190 Y A TATGCC I 1190 L E E CTCTGA I 1280 K P PAAACCC I 1460 L V CTGGTC I 1550 V A GGAAGA I 1640	D NGACGO	1020  / L  TTCTGG  11110  4 G  RGGANC  1290  / L  RGCANC  1380  / L  RGCANC  1470  K  L  RAATTAI  1560  M  Q  RTGCANC  RGANC  11740	V R GTGCG: N S AACAG D T GACACC A D T GACACC A D T GACACC C T ATCGCC V T AC C T T T T T T T T T T T T T T T T	T R CACGCGC 1210  L L CCTGCTG 1300  I A TATCGCT 1390  K Y GAAGTAT 1480  A E CGCGGAA 1570  K Y RAAGTAC 1660  N I TACCACT 1750	N N PARACE R R GCGCG R F T T T T T T T T T T T T T T T T T T	1040 D R ATCGT 1130 D A ACGCT 1220 V S CCGTC 1310 A V CCGTC 14400 F S CCGCG 1580 P A CCGCG 1580	W V TGGGT D GGACGG R C CGCTG E M AGCCA K P AAAACC V CGTCG	OF COATTO	E L TAGGETC  F I TITATC  V E S GAATCC  N T AACACC  S D AGTGAT  P P C C G C C C C C C C C C C C C C C	N S AATTCA 1150 R S CGTAGC 1240 Q Q CAGCAG 1330 T I ACTATT 1420 E G GAAGGC 1510 S K AGTAAG 1600 TCAAAAT 1690 A M CCCCGAT 1780	D GATAIN N N AACC CAGG GGCG L CTCA Q CAGC TT	1070 S L GCCTI 11160 L Q TGCAA 1250 A F CTTTI 1340 R V GGCGTI 1430 G E GCGAA 1520 T T GCCAC 1610 R K GCAA 1700 D I 17700	PRL PCCGCGCCTC 1170  EAR GGAAGCGCGC 1260  FEQ FTTTGAACA- 1350  TTQ PACTACGCA- 1440  ASS AGCATCGTC 1530  MLS CATGCTATCC 1620  QLV GCAGCTGGT 1710  FVT TCGTTACCGGT 1800  IAD	297 327 357 357 387 417 447
E P GAGCC.  K I I AAGAT.  W L TGGCT.  G E GGTGA.  K Y AAATA.  T A ACGGC.  D Q GATCA.	1000 E ACACCC CAACCC T720	Y V ATGTC.  Q Q Q AGCAG K S AGAGC F M TTATG H S ACAGT R A GCGCG I M GGATAA	I PATTCCT I PATTCCT I PATTCCT I 190 Y A TATGCC I 1190 L E E CTCTGA I 1280 K P PAAACCC I 1460 L V CTGGTC I 1550 V A GGAAGA I 1640	D NGACGO	1020  / L  TTCTGG  11110  4 G  RGGANC  1290  / L  RGCANC  1380  / L  RGCANC  1470  K  L  RAATTAI  1560  M  Q  RTGCANC  RGANC  11740	V R GTGCG: N S AACAG D T GACACC A D T GACACC A D T GACACC C T ATCGCC V T AC C T T T T T T T T T T T T T T T T	T R CACGCGC 1210  L L CCTGCTG 1300  I A TATCGCT 1390  K Y GAAGTAT 1480  A E CGCGGAA 1570  K Y RAAGTAC 1660  N I TACCACT 1750	N N PARACE R R GCGCG R P T T T T T T T T T T T T T T T T T T	1040 D R ATCGT 1130 D A ACGCT 1220 V S CCGTC 1310 A V CCGTC 14400 F S CCGCG 1580 P A CCGCG 1580	W V TGGGT D GGACGG R C CGCTG E M AGCCA K P AAAACC V CGTCG	OF COATTO	E L TAGGETC  F I TITATC  V E S GAATCC  N T AACACC  S D AGTGAT  P P C C G C C C C C C C C C C C C C C	N S AATTCA 1150 R S CGTAGC 1240 Q Q CAGCAG 1330 T I ACTATT 1420 E G GAAGGC 1510 S K AGTAAG 1600 TCAAAAT 1690 A M CCCCGAT 1780	D GATAIN N N AACC CAGG GGCG L CTCA Q CAGC TT	1070 S L GCCTI 11160 L Q TGCAA 1250 A F CTTTI 1340 R V GGCGTI 1430 G E GCGAA 1520 T T GCCAC 1610 R K GCAA 1700 D I 17700	PRL PCCGCCCTCC 1177  EAR GGAAGCGCG 1260  FEQTTTGAACA 1350  TTQ TACTACGCAA 1440  ASS AGCATCGTCC 1530  MLS CATGCTATCC 1620  QLV GCAGCTGGTT 1710  FVT TCGGTTACCG	297 327 357 357 387 417 447 477
E P GAGCC.  K I I AAGAT.  W L TGGCT.  G E GGTGA.  K Y AAATA.  T A ACGGC.  D Q GATCA.	1000 E ACACCC CAACCC T720	Y V ATGTC.  Q Q Q AGCAG K S AGAGC F M TTATG H S ACAGT R A GCGCG I M GGATAA	I PATTCCT I PATTCCT I PATTCCT I 190 Y A TATGCC I 1190 L E E CTCTGA I 1280 K P PAAACCC I 1460 L V CTGGTC I 1550 V A GGAAGA I 1640	D NGACGO	1020  / L  TTCTGG  11110  4 G  RGGANC  1290  / L  RGCANC  1380  / L  RGCANC  1470  K  L  RAATTAI  1560  M  Q  RTGCANC  RGANC  11740	V R GTGCG: N S AACAG D T GACACC A D T GACACC A D T GACACC C T ATCGCC V T AC C T T T T T T T T T T T T T T T T	T R CACGCGC 1210  L L CCTGCTG 1300  I A TATCGCT 1390  K Y GAAGTAT 1480  A E CGCGGAA 1570  K Y RAAGTAC 1660  N I TACCACT 1750	N N AAATG  R R GCGCG  C CAGGG  FTTTCT  N AAACC  R CCGAGG  C CGAGG  C CGAGGA	1040 D R ATCGT 1130 D A ACGCT 1220 V S CCGTC 1310 A V CCGTC 14400 F S CCGCG 1580 P A CCGCG 1580	W V TGGGT D GGACGG R C CGCTG E M AGCCA K P AAAACC V CGTCG	OF COATTO	E L TAGGETC  F I TITATC  V E S GAATCC  N T AACACC  S D AGTGAT  P P C C G C C C C C C C C C C C C C C	N S AATTCA 1150 R S CGTAGC 1240 Q Q CAGCAG 1330 T I ACTATT 1420 E G GAAGGC 1510 S K AGTAAG 1600 TCAAAAT 1690 A M CCCCGAT 1780	D GATAIN N AACC CAGG S TCAC CTCA CCAGC CAGC T T GACG	1070 S L GCCTI 11160 L Q TGCAA 1250 A F CTTTI 1340 R V GGCGTI 1430 G E GCGAA 1520 T T GCCAC 1610 R K GCAA 1700 D I 17700	PRL PCCGCGCCTC  EAR GGAAGCGCG 1260  FEQ PTTTGAACA 1350  TTQ PACTACGCAA 1440  ASS AGCATCGTC 1530  MLS CATGCTATC 1620  QLV GGCAGCTGGT 1711  FVT PCGTTACCG 1800	297 327 357 357 387 417 447 477
E P GAGCC.  K I I AAGAT.  W L TGGCT.  G E GGTGA.  K Y AAATA.  T A ACGGC.  D Q GATCA.	1000 E AGARTC 1090 N CAATC 1180 I CATC 1540 C 1630 C CAACC 1720 C C CAACC 1720 C C CAACC 1720 C C CAACC 1720 C C C C C C C C C C C C C C C C C C C	Y V ATGTC.  Q Q Q AGCAG K S AGAGC F M TTATG H S ACAGT R A GCGCG I M GGATAA	I PATTCCT I PATTCCT I PATTCCT I 190 Y A TATGCC 1190 L E CTGGAG 1370 P R CCACGC 1460 L V CTGGTT 1550 V A GGGAAGF 1730 E (GGAGGC 1820	D Victorial Control of the Control o	1020  / L  TTCTGG 11110  4 G  ROGGT/1 1200  R N  GCAACC 1290  / L  ROCTGG 1380  / L  ROCTGG 1380  M Q  ATGCAG 1740  M Q  ATGCAG 1750  M Q  ATGCAG 1750  M Q  ATGCAG 1750  M Q  ATGCAG 1750  M Q  ATGCAG 1830	V R GTGCG: N S AACAG D T GAGCAC A D GGGGGA: I A ATCGCC V A GTTGC: CAGGAC CAGCAC CAGGAC CAGCAC CAGCAC CAGCAC CAGCAC CAGCAC CAGCAC CAGCAC CAGCAC CAGCAC CACCAC CACCAC CACCAC CACCAC CACCAC CACCAC	T R CACGCGC 1210  L L CCTGCTG 1300  I A TATCGCT 1390  K Y GAAGTAT 1480  A E CGCGGAA 1570  K Y TAAGTAC 1660  N I T A CATTAC 1750  I N C GCTCAACCA 1840	N AAATG R R GCGCGCG Q Q CAGGG F TTTCT N AAACC R GCGAGG CAGGAG	1040 D R ATCGT 1130 D A ACGCT 1220 V S TC1310 A V CCGTC 1400 F S CCGGC 1580 E S H N CACAG 17 I TACAG	W V TGGGT  D G GACGG  R C C GCTG  E M AGCCA  K P AAACC  L S TTATC  V C GTCG  V TGTGT	1050 VGGTTM 1140 1140 1140 11410 114	E L TAGGETC  F I FITTATC  V E S GAATCC  N T AACACC  S D AGTGAT  P P P CGGCGG  T F AACCC  V E AGTGGA	1060  N S AATTCA 1150  R S CGTAGC 1240  Q Q CAGCAG 1330  T I ACTATT 1420  E G GAAGGC 1510  S K AGTAAG 1600  TCAAAAT 1690  A M CGCGAT 1780  GAAGGT	D GATA	1070 S L GCCTT 1160 L Q TGCAG 1250 A F 1340 R V GCGTT11 1340 T T CCAGC 1610 R R R K 1700 D I GATTT 13ATT 13A	PRL PCCGCCCTCC 1177  EAR GGAAGCGCG 1260  FEQTTTGAACA 1350  TTQ TACTACGCAA 1440  ASS AGCATCGTCC 1530  MLS CATGCTATCC 1620  QLV GCAGCTGGTT 1710  FVT TCGGTTACCG	297 327 357 357 387 417 447 477
E P GAGCC.  K I I AAGAT.  W L TGGCT.  K Y AAAATA.  T A ACGGC.  \$ TGACC.  K TGACC.	1000 E ACACCC 1720 C ACCC 1720 C ACACCC 1720 C ACCC 17	Y V ATGTC.  Q Q Q AGCAG K S AGAGC F M TTATG H S ACAGT I M GCGCG GATAA K L AAACT	1010 I P ATTCTC 11100 Y A TATGCC 11190 L E CTGGAG 1280 K P AAACCC 1370 P R ACCCACGC 1460 L V CTGGTC 1640 V A GTGGCA 1730 E GGAAGA 1730 E GGAAGA 1730	D CGACG	1020  / L  TTCTGG  11100  4 G  TGGGT/ 1200  R N  G1290  / L  TGCTGG  1380  / L  TGCTGG  TG	V R STECCS: N S S AACAGG D T GACACG A D GCGGA*  I A ATCGC V A T GCTCAM	TO A CATTACT TO TO A CATTACT TO	N PARCE N PARC	1040 D R ATCGT 1130 D A ACGCT 1220 V S CCGTC 1490 P A CCGCG 1490 P A CCGCG 1670 H N TACAT 1760 Y I TACAT 1760 O D O O O O O O O O O O O O O O O O O O	W V V V CGTCG	1050  Y GTTM GGTTM CCCAG 1230  I I 1320  I 1320  I 1320  I 1500  L ATTGG 1590  CATTG 1680  L ATTGG L AGATT 1770  L KATTGAG 1860  Y A	E L SAGCTC F I FTTATC V E FCGAA E S GAATCC N T T AAACACC V F AAGTGAT	1060  N S AATTCA 1150  R S CGTAGC 1240  Q Q CAGCAG 1330  T I ACTATT 1420  S K AGTAAG 1600  S N TCAAAAT 1690  X N TCAAAAT 1780  K V GAAGGT 1780  K V GAAGGT 1870	D GATA' N AACC CAGG CAGG STCAC GGCG CTCA CTCA GCGCG CAGC GCGCG GCGCG GCGCG GCGCG GACG	1070 S L 1 1160 L Q TGCAG 1250 A F T 1340 R V GCGTT1 1340 G E GCGAA 1520 T T T CCCAAA 1700 G B K GTAAC 1700 C CAAA 1780 Q : CAAAA 1880	PRL PCCGCCCTC 1177  EAR GGAAGCGCG 1260  FEQTTTGAACA 1350  TTQ FACTACGCA 1444  ASS AGCATCGTC 1534  ALS CATGCTATCC 1620  QLV GCAGCTGGT 17710  FVT FCGTTACCG 1890	297 327 357 357 367 417 50 447 447
E P GAGCC.  K I I AAGAT.  W L TGGCT.  K Y AAAATA.  T A ACGGC.  \$ TGACC.  K TGACC.	1000 E ACACCC 1720 C ACCC 1720 C ACACCC 1720 C ACCC 17	Y V ATGTC.  Q Q Q AGCAG K S AGAGC F M TTATG H S ACAGT I M GCGCG GATAA K L AAACT	1010 I P ATTCTC 11100 Y A TATGCC 11190 L E CTGGAG 1280 K P AAACCC 1370 P R ACCCACGC 1460 L V CTGGTC 1640 V A GTGGCA 1730 E GGAAGA 1730 E GGAAGA 1730	D CGACG	1020  / L  TTCTGG  11100  4 G  TGGGT/ 1200  R N  G1290  / L  TGCTGG  1380  / L  TGCTGG  TG	V R STECCS: N S S AACAGG D T GACACG A D GCGGA*  I A ATCGC V A T GCTCAM	TO A CATTACT TO TO A CATTACT TO	N PARCE N PARC	1040 D R ATCGT 1130 D A ACGCT 1220 V S CCGTC 1490 P A CCGCG 1490 P A CCGCG 1670 H N TACAT 1760 Y I TACAT 1760 O D O O O O O O O O O O O O O O O O O O	W V V V CGTCG	1050  Y GTTM GGTTM CCCAG 1230  I I 1320  I 1320  I 1320  I 1500  L ATTGG 1590  CATTG 1680  L ATTGG L AGATT 1770  L KATTGAG 1860  Y A	E L SAGCTC F I FTTATC V E FCGAA E S GAATCC N T T AAACACC V F AAGTGAT	1060  N S AATTCA 1150  R S CGTAGC 1240  Q Q CAGCAG 1330  T I ACTATT 1420  S K AGTAAG 1600  S N TCAAAAT 1690  X N TCAAAAT 1780  K V GAAGGT 1780  K V GAAGGT 1870	D GATAN N AACC Q CAGGG GGCG L CTCA GGGCG R GGCGC GGCGC GGCGC GGCGC GGCGC GGCGC GGCTT GACGG GGCTT GGCGT GGCGT GGCGT GGCGT GGCGC GGCTT GGCTT GGCGC GGCTT GGCCTT GGCGC GGCTT GGCTT GGCGC GGCTT GGCTT GGCTT GGCGC GGCTT GGCGC GGCT GGCTT GGCGC GGCT GGCTT GGCGC GGCT GGCTT GGCGC GGCTT GGCGC GGCT GGCTT GGCGC GGCT GGCTT GGCGC GGCT GGCTT GGCGC GGCT GGCGC GGCT GGCTT GGCGC GGCT GGCTT GGCGC GGCT GGCT GGCT GGCT GGCT GGCGC GGCT GGCGC GGCT GGCT GGCGC GGCT GGCGC GGCT GGCGC GGCT GGCT GGCGC GGCT GGCT GGCGC GGCT GGCC GGCT GGCGC GGCC GGCT GGCGC GGCT GGCT GGCGC GGCT GGCT GGCT GGCT GGCT GGCGC GGCT	1070 S L 1 1160 L Q TGCAG 1250 A F T 1340 R V GCGTT1 1340 G E GCGAA 1520 T T T CCCAAA 1700 G B K GTAAC 1700 C CAAA 1780 Q : CAAAA 1880	PRL PCCGCGCCTC 1170 EAR GGAAGCGCGC 1260 FEQ FTTTGAACA 1350 TTQ FACTACGCA 1440 ASS AGCATCGTC 1530 MLS CATGCTATC 1620 OLV GCAGCTGGT 1710 FVT TCGTTACCG 1890 IAD TTGCGGACG	297 327 357 357 387 417 447 50 447

restriction at the unique <u>Hind</u>III site and subjected to <u>Bal</u>31 digestion in order to remove approximately 200 bp from the <u>ClaI</u> fragment. Suitably sized deletions were selected by screening plasmid digests and the endpoints of these deletions were subsequently determined after cloning in mp8. Two such deletions pMM31 and pMM32 terminated at positions 1759 and 1776 and had therefore lost 196 bp and 179 bp respectively of the <u>ClaI</u> fragment. These deleted plasmids still synthesised a 75 kDal polypeptide when used as templates in an S-30 <u>in vitro</u> transcription/translation system (Fig. 3).

The predicted amino acid composition of NtrA is given in Table 1. The protein has a significantly greater than average proportion of acidic and acid amide residues which is consistent with previous studies which estimated the pI of NtrA as <5.0 (7).

The Analyseq program of Staden (29) was used to identify a potential transcription initiation site for <a href="https://ntranscription.com/ntransc

Analysis of the sequence downstream of <a href="https://ntranslike.com/ntranslike

Fig. 2. Nucleotide sequence of the 1966 bp EcoRI-HindIII fragment from pMM17 (see Fig. 1). Nucleotides are numbered from the first base of the EcoRI site. Proposed ribosome binding sites (positions 270, 1725) and ntrA promoter sequences (positions 180 and 206) are underlined. The sequence is extended 14 bp beyond the HindIII site to indicate premature termination of ORF2 in the vector sequences. Numbers at the right hand side of the figure indicate numbers of the amino acid residues in NtrA.

Table 1. Amino acid composition (mole '%) of NtrA compared with that of other proteins.

	K. pneumoniae	E. COLI	E. COI1	B. subtilis	
Amino acids	<u>ntrA</u>	rpoD	nusA	rpoD	protein
		(ref. 30)	(ref. 32)	(ref. 31)	(ref. 48)
Acidic (D+E)	16.7	20.4	18.6	21.1	11.5
Acid + acid amide (D+N+E+Q	27.6	28.4	26.1	28.0	19.8
Basic (K+R+H)	11.2	14.5	12.3	16.0	13.5
Hydrophobic (L+V+I+M)	26.7	25.4	27.0	24.9	20.2
Aromatic (F+Y+W)	5.4	5.2	5.0	6.0	8.3
Charged (D+E+K+R+H)	27.9	34.9	30.9	37.1	25.1
Aliphatic (A+G)	10.9	12.1	17.6	11.3	16.9
Hydroxyl (S+T)	13.2	10.8	8.9	9.7	13.1
Molecular weight (kDal)	*				
(i) from DNA sequence	54		70	55	43
(ii) SDS-PAGE	75	82	2-90	69	55
pI	<5.0	4.8	-5.1	4.6	-

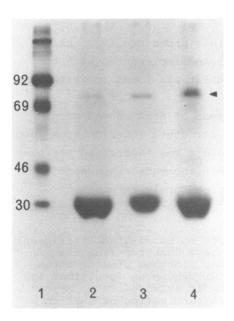


Fig. 3. SDS-polyacrylamide electrophoresis of in vitro transcription translation products synthesised in an E. coli S-30 system. Track 1 - [1 $\frac{1}{4}$ C] mol.wt. markers, Track 2 - pMM17, Track 3-pMM32, Track 4 - pMM31. The arrow indicates the ntrA product of 75 kDal. The  $\frac{30}{30}$  kDal polypeptide is the bla product from the pBR327 vector.

sequence at position 1725 and is potentially a second gene in the ntrA operon.

### DISCUSSION

# The nucleotide sequence of ntrA

The expression of the vector-encoded tetracycline resistance (tet) gene in two independently constructed <a href="https://doi.org/nc.10">ntrA</a> plasmids pMM17 (7) and pFB71 (6) was originally attributed to a fortuitous transcriptional fusion of tet to the <a href="https://doi.org/nc.10">ntrA</a> promoter. Our sequence supports this proposition and indicates that transcription initiated at the <a href="https://doi.org/nc.10">ntrA</a> promoter could continue through <a href="https://doi.org/nc.10">ntrA</a>, through the prematurely terminated second <a href="https://doi.org/nc.10">org/nc.10</a> and into the tet gene.

The identification of a second <u>orf</u> apparently in the same operon as <u>ntrA</u> raises the question of the function of this second gene. If these genes are co-transcribed it is possible that the functions of the two gene products are related. Relatively little genetic analysis of the <u>ntrA</u> region has been undertaken (10,13) and further experiments to determine the

Ec RpoD	(111)	LTREGEIDIAKRIED-G * AKKEMVEANLRLVISIAKKYTNR
Bs RpoD	(114)	LSAKEEIAYAQKIEE-GDEESKRRLAEANLRLVVSIAKRYVGR
Bs SpoIIG	(39)	LSKDEEQVLLMKLPN-GDQAARAILIERNLRLVVYIARKFENT
Ec HtpR	(29)	LSADEERALAEKLHYHGDLEAAKTLILSHLRFVVHIARNYAGY
Ec RpoD Bs RpoD Bs SpoIIG Ec HtpR		GLQFLDLIQEGNIGLMKAVDKFEYRKGYKFSTYATWWIRQAI GMLFLDLIHEGNMGLMKAVEKFDYRRGYKFSTYATWWIRQAI GINIEDLISIGTIGLIKAVNTFNPEKKIKLATYASRCIENEI GLPQADLIQEGNIGLMKAVRRFNPEVGVRLVSFAVHWIKAEI

Fig. 4. Alignment of homologous regions in E. coli RpoD and HtpR, and B. subtilis RpoD and SpoIIG. Amino acid residues are defined by the standard one letter code. Figures in brackets refer to the position of the first amino acid shown in each protein. The \* in E. coli RpoD indicates the position of a 248 residue insert not present in the other sequences. Residues conserved in two or more proteins are underlined and I, L and V are considered as conservative substitutions.

phenotype of mutations in this region will be of interest. Comparison of the NtrA sequence with other sigma proteins

A number of genes which encode known RNA polymerase sigma factors have been cloned and sequenced. These include E. coli rpoD (30) and htpR (34,35), B. subtilis rpoD (31) and spoIIG (36) and B. subtilis bacteriophage SPO1 genes 28 (gp28) (37) and 34 (gp34) (38). The sequence is also available for E. coli NusA which, although not a sigma factor, is known to interact with RNA polymerase thereby modifying both transcription termination and antitermination (32). Previous authors have compared various of these sequences and some regions of homology have been identified. E. coli RpoD and B. subtilis RpoD are highly homologous (31) and a region of 84 residues is conserved between these two proteins and E. coli HtpR and B. subtilis SpoIIG (31,34,36) (see Fig. 4). This homology is interrupted in E. coli RpoD by an insert of 248 residues not present in the other sequences (31) and consequently the conservation of the first twenty residues shown in Fig. 4 was not recognised in previous comparisons (34,36). No amino acid sequences homologous to this conserved region are present in Gp28, Gp34, or NusA and no comparable homology could be identified in NtrA.

By further sequence analysis we have now identified two other regions of homology present in  $\underline{\text{E. coli}}$  RpoD, HtpR, NusA;  $\underline{\text{B. subtilis}}$  RpoD, SpoIIG and SPOI Gp34 (Fig. 5A). These regions

			1	10	20	30
Α.	Ec RpoD	(472) (565)		GREPTPEELAE NTDYTLEEVGK		
	Bs RpoD	(231) (324)	<b>₫</b> bro	GREPTPEELAE GRTRTLEEVGK	DMOLTPEKV	REILKIAQ
	Ec HtpR	(135) (245)	Lefi	NLRKTKORLOW ONKSTLOELAD	ifn@devem <mark>v</mark> #	RELGVIS
	Ec NusA	(372)	LVE	EGFSTLEELAY RGVOTLEDLAE	VPMKELLEIE	CCLDEPTV
	Bs SpoIIG	(158) (198)	Grdi	DDIITKDIEAN EEEKTOKDVAL	IVDKKLLKKAT	EQLNERE
	SPO1 Gp34	(89) (148)	LKR	INGETSLYVK IRKKTLQELA	EDGEVLELO	AHMDATT
	Kp NtrA	(332) (379)		SRND <u>TL</u> LR <u>V</u> -S SRVT <u>TQ</u> KY <u>L</u> HS		
В.	GalR TetR LacR P22Cro		1	MATIKD <u>VA</u> F IEGUTTRK <u>LA</u> C MKPVTLYDVAF DHFGTORAVAK	YAGVSYQIVS	WHVKNKR SRVVNQAS
	Fnr Cap		REFI	RLTMTRGD <u>IG</u> DIKITRQ <u>EIG</u> C	YLGITVETIS	RLLGFRQ

Fig. 5. A. Homologous regions in E. coli RpoD, HtpR, NusA; B. subtilis RpoD, SpoIIG, SPOI Gp34 and K. pneumoniae NtrA. Amino acid residues are defined by the standard one letter code and conserved residues are underlined (I, L and V are considered as conservative substitutions). Figures in brackets refer to the position in the protein of the first amino acid in each sequence. Figures above the sequences are for reference purposes (see text). B. Comparative sequences from known site-specific DNA binding proteins aligned according to ref. 39.

are of particular interest as they include features which have been proposed to be characteristic of site-specific DNA binding proteins (39). In all cases these homologous regions are at the C-terminal end of the protein and are separated by a region of 40 to 100 residues. Two of these regions, residues 331 to 352 in B. subtilis RpoD and residues 253 to 272 in E. coli HtpR, have previously been identified as possible DNA-binding domains (31,34).

The homologous regions have a number of features in common which identify them as potential DNA binding sites (Fig. 5A).

(i) Conservation of hydrophobic residues at positions 1, 12, 18, 23 and 26; (ii) an invariant threonine at position 8; (iii) conserved polar residues at positions 10 and 11; (iv) a

	10	10
Ec RpoD Bs RpoD Ec HtpR Bs SpoIIG SPO1 Gp34	(479) <u>TPEELAERMLM</u> (238) <u>TPEEIAEDMDL</u> (142) <u>TKQRLG</u> WFNQD (165) <u>TKDIEANVDKK</u> (96) <u>TSLYVKNEDGE</u>	(572) TLEEVGKQFDV (331) TLEEVGKVFGV (252) TLQELADRVGV (205) TQKDVADMMGI (155) TLQELAQEEGV
Ec NusA	(379) <u>TLEELAYVPMK</u>	(454) <u>TLEDLAEQ</u> GID
Kp NtrA	(339) <u>TLLRV</u> -SRC <u>IV</u>	(386) TQKYLHSPRGI

Fig. 6. Selected sequences from Fig. 5 aligned to show the higher degree of homology present in the more C-terminal of the two homologous regions in each protein. Other details are as in Fig. 5.

conserved alanine or glycine at position 13. For comparative purposes the relevant sequences from a number of known sitespecific DNA binding proteins are shown in Fig. 5B.

When intergenic comparisons are made for the two regions in each protein the most C-terminal of the two regions shows the greatest degree of conservation (Fig.6), particularly with respect to the conserved glycine and adjacent hydrophobic residue at positions 17 and 18. These two residues are characteristic of the tight turn between the two  $\alpha$ -helices involved in DNA binding but neither is apparently invariant (39).

When the NtrA sequence was searched for comparable sequences to those described above, two potential regions (residues 332 to 361 and 379 to 409) were identified (Fig. 5A). In each case the degree of homology is less than that seen in the other six sequences. Analysis of the SPO1 Gp28 sequence failed to identify any regions homologous to the sequences in Fig. 5A. Role of potential DNA-binding domains?

The identification of two quite closely linked (40-100 residues apart) potential DNA-binding sites in five functionally related proteins raises the question of the possible functions of such sites. Whilst RNA polymerase core enzyme alone possesses catalytic activity, it is the sigma subunit which is required for promoter recognition. Isolated sigma factor from <u>E. coli</u> and <u>B. subtilis</u> has been shown to bind to supercoiled DNA, but not in a site-specific fashion (40,41). However, cross-linking studies with RNA polymerase holoenzyme have shown that sigma can cross-

link to the promoter (42) suggesting that promoter selection is probably dictated by direct interaction of sigma factor, as part of the holoenzyme complex, with specific nucleotides in the promoter. In <u>E. coli</u> two different sigma factors are known (30,43) and in <u>B. subtilis</u> five different sigma factors have been identified as well as two phage SPO1-encoded sigmas, and each apparently recognises specific -35 and -10 promoter consensus sequences (44).

Two models for the role of sigma factors in promoter selection have been proposed (45). In the 'core-conformation' model each sigma contacts a characteristic -10 sequence and induces a core-conformation that favours a particular canonical sequence at the -35 position. In the direct-contact model, favoured by Losick and Pero (45), sigma contacts both the -35 and -10 regions, either simultaneously or sequentially, during the formation of the RNA polymerase-promoter complex. Such a model requires each sigma factor to have two domains which will mediate -35 and -10 recognition. These domains would not necessarily be expected to resemble closely the structures found in sitespecific DNA binding proteins as the nature both of the DNA sequence recognised and the type of protein DNA interaction are probably different in sigma factors. Nevertheless, the identification of two domains which resemble the consensus DNA binding site of site-specific DNA binding proteins suggests that these regions may be those which are concerned with recognition of the -35 and -10 promoter sequences. An alternative, but perhaps less likely, role for these conserved sequences is that of a protein/protein recognition domain, i.e. a region of interaction with one or more subunits of RNA polymerase core enzyme.

As described earlier, <a href="https://ntranslive.com/ntm2">ntr-activatable</a> promoters have an entirely different consensus, both in sequence and spacing, from that found in other prokaryotic promoters. Hence, if NtrA is indeed a sigma factor it might be expected to differ from other sigma factors in respect of any potential DNA-recognition domains. The two regions we have identified in NtrA are similar but not distinctly homologous to the very well conserved regions in other sigma factors and like the other regions they are at the C-

terminus of the protein. It remains to be demonstrated whether these NtrA sequences or those identified in the other sigma factors do indeed play a direct role in promoter selection.

The role of the consensus sequences in NusA may differ from that in other sigma factors. However, a non-symmetrical DNA sequence (boxA) has been identified as a potential recognition site for NusA suggesting that transcription termination by an RNA polymerase/NusA complex requires interaction between NusA and the boxA sequence (46). Recent studies of transcriptional autoregulation by NusA suggested that a NusA: chloramphenicol transacetylase (CAT) hybrid protein, in which CAT is fused after residue 456 of NusA, is defective in normal regulation (47) indicating that integrity of the second consensus sequence (residues 447-477) may be necessary for normal NusA function. is notable that the two NusA DNA-recognition sequences show much greater inter-sequence homology than that seen between other intergenic pairs and in NusA the region of homology extends over 48 residues of which 19 are identical and 6 show conservative replacements.

The absence of any obvious homology between SPO1 Gp28 and other sigma factors is surprising. However until the nature of the interaction between sigma factors and RNA polymerase core enzyme, and between RNA polymerase and promoter sequences, is better understood it may not be possible to predict potentially homologous domains from the primary amino acid sequence of sigma factors.

In this paper we believe we have identified some potentially interesting sequences in a number of sigma factors and in NtrA. It remains for future studies to determine whether these sequences are functionally homologous and structurally significant in the mode of action of RNA polymerase and associated sigma factors.

## **ACKNOWLEDGEMENTS**

We would like to thank Ray Dixon, Martin Drummond and Martin Buck for much advice and valuable discussion. We also thank Brenda Hall for typing the manuscript.

Since this paper was submitted for publication a comparable analysis of bacterial sigma factors has been published by Stragier et al., (FEBS Letts. 187, 11-15, 1985). These authors analyse the E. coli RpoD and HtpR and B. subtilis RpoD and SpoIIG sequences and identify the same potential DNA binding sites as those described in this paper.

#### REFERENCES

- Leonardo, J.M. and Goldberg, R.B. (1980) J.Bacteriol., 142, 99-110.
- Espin, G., Alvarez-Morales, A. and Merrick, M. (1981) Mol. 2. Gen.Genet., 184, 213-217.
- 3. Espin, G., Alvarez-Morales, A., Cannon, F., Dixon, R. and Merrick, M. (1982) Mol.Gen.Genet., 186, 518-524. Merrick, M. (1983) EMBO J., 2, 39-44.
- 4.
- 5. De Bruijn, F.J. and Ausubel, F.M. (1981) Mol.Gen.Genet., 183, 289-297.
- 6. De Bruijn, F.J. and Ausubel, F.M. (1983) Mol.Gen.Genet., 192, 342-353.
- 7. Merrick, M.J. and Stewart, W.D.P. (1985) Gene, 35, 297-303.
- Alvarez-Morales, A., Dixon, R. and Merrick, M. (1984) EMBO J. 8. 3, 501-507.
- 9. Dixon, R. (1984) Nucleic Acids Res., 12, 7811-7830.
- Garcia, E., Bancroft, S., Rhee, S.G. and Kustu, S. (1977) 10. Proc.Natl.Acad.Sci.USA, 74, 1662-1666.
- McFarland, N., McCarter, L., Artz, S. and Kustu, S. (1981) Proc.Natl.Acad.Sci.USA, 78, 2135-2139. 11.
- Pahel, G., Rothstein, D.M. and Magasanik, B. (1982) 12. J.Bacteriol., 150, 202-213. Magasanik, B. (1982) Ann.Rev.Genet., 16, 135-168.
- 13.
- Rothman, N., Rothstein, D., Foor, F. and Magasanik, B. (1982) 14. J.Bacteriol., 150, 665-671.
- 15. Beynon, J., Cannon, M., Buchanan-Wollaston, V. and Cannon, F. (1983) Nature, 301, 302-307.
- 16. Drummond, M., Clements, J., Merrick, M.J. and Dixon, R. (1983) Nature, 301, 302-307.
- 17. Ow, D.W., Sundaresan, V., Rothstein, D.M., Brown, S.E. and Ausubel, F.M. (1983) Proc.Natl.Acad.Sci.USA, 80, 2524-2528.
- 18.
- Dixon, R.A. (1984) J.Gen.Microbiol., 130, 2745-2755. Hawley, D.K. and McClure, W.R. (1983) Nucleic Acids Res., 19. 11, 2237-2255.
- Ausubel, F.M. (1984) Cell, 37, 5-6. 20.
- Alvarez-Morales, A. and Hennecke, H. (1985) Mol.Gen.Genet., 21. 199, 306-314.
- Brigle, K.E., Newton, W.E. and Dean, D.R. (1985) Gene, 22. in press.
- 23. Robson, R., Jones, R., Woodley, P. and Evans, D. (1985) in Proceedings of 6th International Symposium on Nitrogen Fixation, Evans, H.J. and Newton, W.E. Eds., Nijhoff/Junk, Boston (in press).
- 24. Dixon, R. (personal communication).
- Messing, J. (1983) in Methods in Enzymology, Wu, R., 25. Grossman, L. and Moldave, K. Eds., Vol. 101, pp.20-77. Academic Press, New York.
- 26. Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) Proc.Natl. Acad.Sci.USA, 80, 3963-3965.

- Casadaban, M.J., Chou, J. and Cohen, S.N. (1980) J.Bacteriol. 143, 971-980.
- 28. Staden, R. (1982) Nucleic Acids Res., 10, 4731-4751.
- 29. Staden, R. (1984) Nucleic Acids Res., 12, 521-538.
- Burton, Z., Burgess, R.R., Lin, J., Moore, D., Holder, S. and Gross, C.A. (1981) Nucleic Acids Res., 9, 2889-2903.
- Gitt, M.A., Wang, L-F. and Doi, R.H. (1985) J.Biol.Chem., 260, 7178-7185.
- 32. Ishii, S., Ihara, M., Maekawa, T., Nakamura, Y., Uchida, H. and Imamoto, F. (1984) Nucleic Acids Res., 12, 3333-3342.
- Castano, I. and Bastarrachea, F. (1984) Mol.Gen.Genet., 195, 228-233.
- 34. Landick, R., Vaughn, V., Lau, E.T., VanBogelen, R.A., Erickson, J.W. and Neidhardt, F.C. (1984) Cell, 38, 175-182.
- 35. Yura, T., Tobe, T., Ito, K. and Osawa, T. (1984) Proc.Natl. Acad.Sci.USA, 81, 6803-6807.
- Stragier, P., Bouvier, J., Bonamy, C. and Szulmajster, J. (1984) Nature, 312, 376-378.
- Costanzo, M. and Pero, J. (1983) Proc.Natl.Acad.Sci.USA, 80, 1236-1240.
- Costanzo, M., Brzustowicz, L., Hannett, N. and Pero, J. (1984) J.Mol.Biol., 180, 533-547.
- Pabo, C.O. and Sauer, R.T. (1984) Ann. Rev. Biochem., 53, 293-321.
- 40. Kudo, T. and Doi, R.H. (1981) J.Biol.Chem., 256, 9778-9781.
- Kudo, T., Jaffe, D. and Doi, R.H. (1981) Mol.Gen.Genet., 181, 63-68.
- 42. Simpson, R.B. (1979) Cell, 18, 277-285.
- Grossman, A.D., Erickson, J.W. and Gross, C.A. (1984) Cell, 38, 383-390.
- Johnson, W.C., Moran, C.P. and Losick, R. (1983) Nature, 302, 800-804.
- 45. Losick, R. and Pero, J. (1981) Cell, 25, 582-584.
- 46. Friedman, D.I. and Olson, E.R. (1983) Cell, 34, 143-149.
- Plumbridge, J.A., Dondon, J., Nakamura, Y. and Grunberg-Manago, M. (1985) Nucleic Acids Res., 13, 3371-3388.
- 48. Dayhoff, M.O., Hunt, L.T. and Hurst-Calderone, S. (1978) in Atlas of Protein Sequence and Structure, Dayhoff, M.O. Ed., Vol. 5, pp.363-369.